The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 160-177

基於雙工音高感知模型之神經網路旋律抽取演算法

The duplex model of pitch perception inspired neural network for melody extraction

周歆, 冀泰石
國立交通大學電機工程學系
chousmit.04g@g2.nctu.edu.tw, tschi@mail.nctu.edu.tw

摘要

本論文根據聽覺的觀點提出利用類神經網路建構旋律抽取的方法，針對複音音樂進行旋律的抽取。根據傳統心理聲學音高分析理論，人在音高的解析分為頻譜模型和時間模型。在此論文中，我們先對個別模型進行探討並建構模型評比效能，觀察個別模型的訓練結果與聽覺理論是否相同，並依據結果建構出頻譜模型上的聽覺模板。再進一步針對頻譜模型上高頻諧音無法解析的缺失利用時間模型補足，建構出雙工模型。由實驗結果可知由時間模型補足頻譜模型無法解析的頻段有助於提升旋律抽取及音高判別。此實驗結果也證明以心理聲學為基礎來建構類神經網路確實可用於音樂資訊檢索的相關應用中。

1. 生理聽覺現象與特性

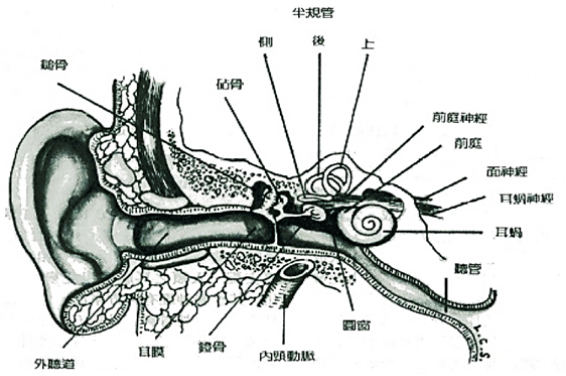

圖 1.1：耳朵基本構造[1]

圖(1.1)為人耳的基本構造，主要包含外耳、中耳、以及內耳三個部份。外耳包含耳殼及外聽道，耳殼負責收集外界聲音，經由外聽道傳至耳膜；中耳由三小聽骨(鎚骨、砧骨及鐙骨)組成；而內耳最重要的部分為耳蝸，耳蝸被基底膜所分成兩層，其內部充滿組織液，因組織液的流動在基底膜上產生行進波(Traveling wave)。依據外界生音的頻率不同，行進波會在基底膜上不同的位置產生最大振振幅，如圖(1.2)所示。

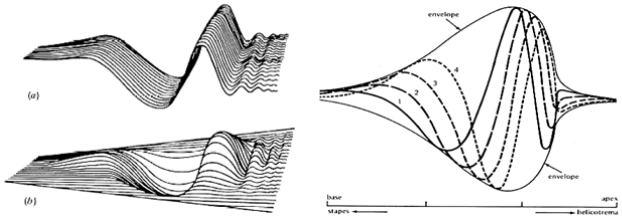

圖 1.2：基底膜上行進波[2]

由於基底膜上的質地和寬度差異，靠近膜底部(前端，base)的質地較硬寬度較窄；而靠近頂部(後端，apex)較寬軟，如圖(1.3)所示。使得不同頻率的聲音，在基底膜上所產生的行進波會在不同的位置為產生最大振幅。因此，基底膜可視為一系列的頻率濾波器。較低頻的聲音會在較遠處才產生最大共振;而頻率較高的聲音,在靠近卵圓窗膜底部的位置就會產生最大共振,此一濾波頻率範圍大約為 20 Hz 到 20000 Hz,即正常人類的聽覺範圍。外界的聲波,經由外耳、中耳、內耳的順序依序傳遞，將聲波轉換成最後的電訊號，使我們能聽到聲音。

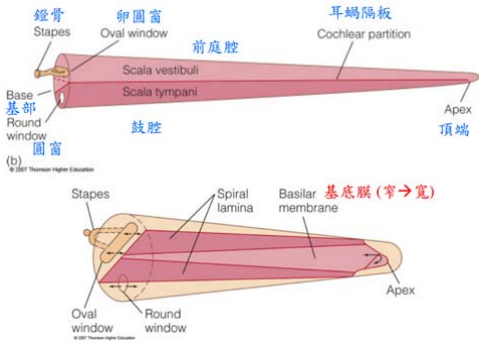

圖 1.3：基底膜構造示意圖[3]

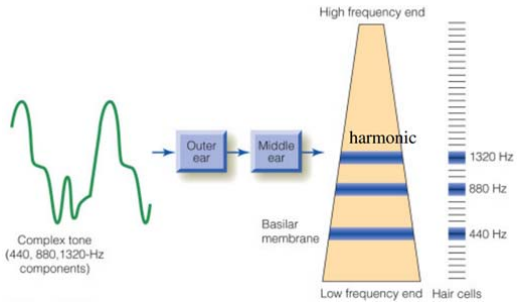

圖1.4：複音在基底膜上的響應分布[3]

上圖(1.4)表示一個複音訊號經由外耳、中耳再到耳蝸內的基底膜上的反應。可看出此複音是由 440、880、1320Hz 三個頻率呈倍數的單音組合，三種頻率分別會在基底膜不同的位置有最大共振效應可視為分頻的效果，由圖可看出在頻率與其共振的位置呈現對數關係。

文獻[4]將人在音高的判別流程分為四個階段如圖(1.5)，前兩步驟為帶通濾波器及半波整流，對應於人的耳蝸構造及毛細胞的放電反映。然而後兩者的週期性偵測及神經彼此間的交互反映尚未發現生理上的證據，但經由一些現象及結果可推測判斷音高可能的機制。

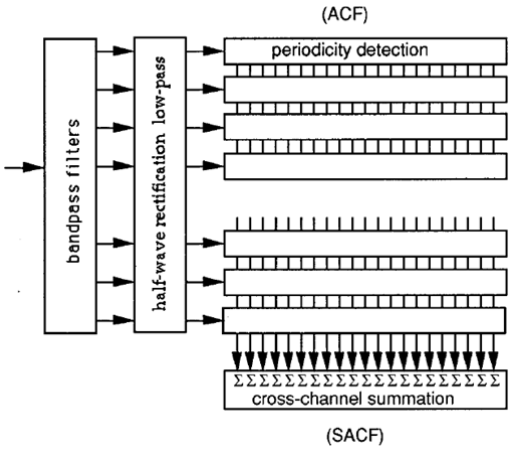

圖1.5 判斷音高的四階段流程圖[4]

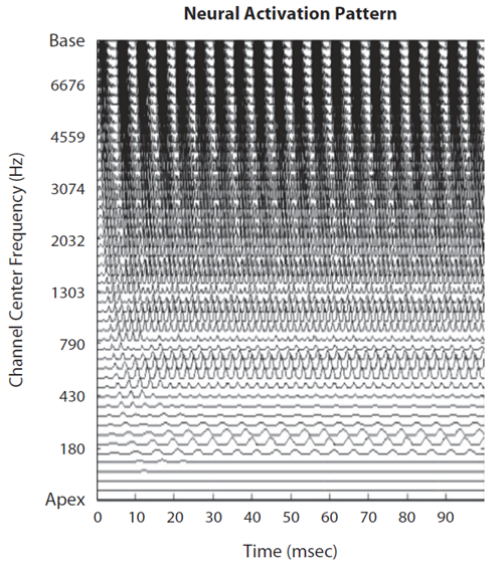

圖 1.6：基頻為 200Hz 的複音在聽神經上的模擬反應[5]

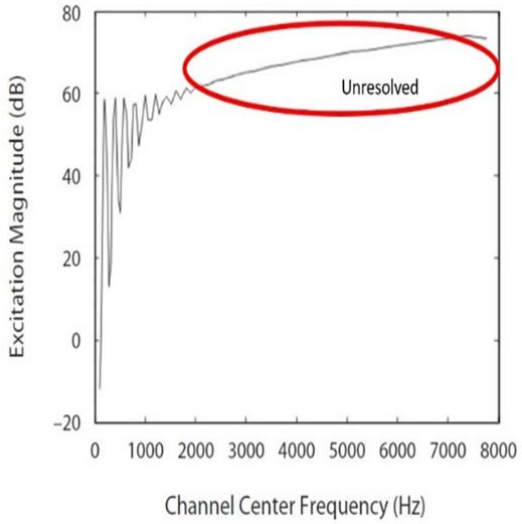

圖 1.7：200Hz 複音刺激聽神經，將反應能量加總後的神經激發反應[5]

聲音在基底膜的分頻在不同地方產生共振後會發送電訊號至腦部，圖為 PATTERSON 和 ALLERHAND [5]模擬當聽到基頻為 200Hz 的複音時神經所產生的激發反應，基底膜的底部到頂端分別對應到高頻到低頻，由圖(1.6)可看出低頻對於諧音的解析較好，越高頻則越差，若將總能量依時間加總起來則會得到圖(1.7)。可發現圖(1.7)中可解析的頻率約至 2000Hz，也就是第 10 個諧音(harmonic)，超過第 10 個的諧音則在此神經激發圖中無法被解析出來。因此，若此時的聲音為 500Hz 的複音，則大約超過 5000Hz 的諧音無法被解析出來，此解析度與耳蝸的濾波器頻寬(critical bandwidth)有關。

基於此現象，頻譜模型(Spectral model)在 1970 年代被提出，藉由解剖學及訊號處理的觀點，聲音訊號經由耳蝸有順序性的分頻可以解析出複音(complex tone)中的諧音成分，而且其諧音的的排列與音高有密切的相關。因此當時的學者認為人類在認知音高上存在一些固定的模板樣式(pattern)，人的生理構造或是腦中有這些模板樣式以致我們可以辨認出不同的音高，基於這個理念提出了許多音高辨識的方法。多數的模型皆是先建立出音高模板(template)，再針對輸入訊號做配對，找出最適合的模板從而決定因高，如圖(1.8)所示。這個方法成功地解釋大部分的生心理聽覺現象，但仍存在一些不能由此模型解釋的盲點，例如殘餘音高的問題(Residual pitch)。

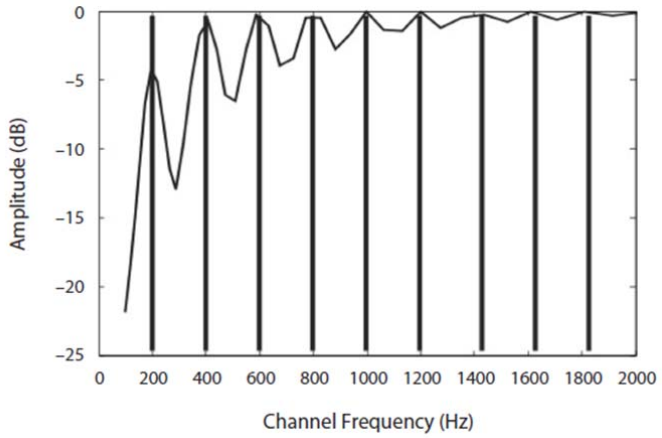

圖 1.8：頻譜模型偵測音高示意圖[6]

然而實際上人仍可聽到高於第 10 個諧音頻律以上的複音結構，這是因為人除了在頻率上解析音高之外，在時間上也可以解析音高，因此有一些學者從其他的觀點來解釋生心理聽覺現象。下圖(1.9)顯示一個間隔 5ms 的脈衝串聲音經對數頻律分佈的濾波器組分頻後在時間軸上的關係。可看出在低頻時因濾波器頻寬較窄，所以可解析出個別的諧音，例如 200、400、600、800Hz，但在高頻的部分因濾波器頻寬很寬無法解析出個別的諧音，但在時間軸上仍可看出明顯的間隔，而此間隔的時間長短也反映了此聲音的音高。

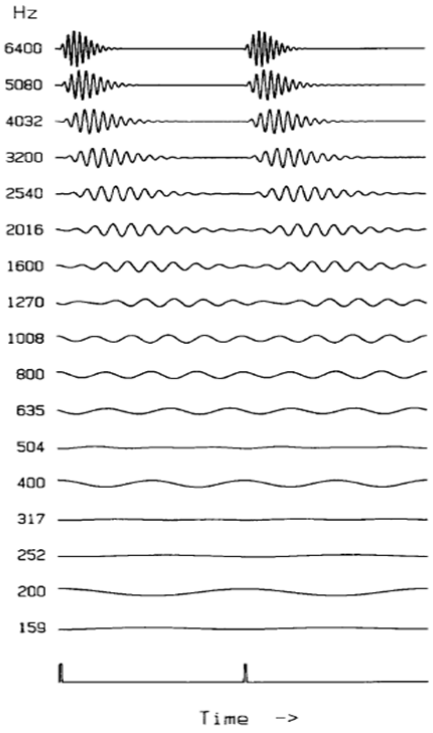

圖 1.9：在時間上不同頻率的響應[1]

因為頻譜模型所存在的一些缺陷，且發現時間上也有包含音高的資訊既而提出時間模型(temporal model)。多數支持時間模型的學者認為人在音高感知判斷上皆是基於自相關函數(autocorrelation)，$A(\tau)$，其數學式如下:

$$A(\tau)=\int x(t)x(t+\tau)dt$$

其中 $x(t)$ 為時間軸的波形訊號，τ 為時間延遲。因訊號包含週期成份，自相關函數在於計算此訊號於其自身在不同時間點的互相關，也就是找出重複模式。標準化自相關函數則是在內積後除上時間延遲。若訊號包含週期成份，計算出的自相關函數也會存在周期性，而這些顯著周期的倒數就是基頻，有可能是人所認知的音高資訊。因為此方法探究時間上的變化資訊，支持時間模型的學者認為此模型可以解釋頻譜模型上高頻諧音無法解析時的音高感知現象。

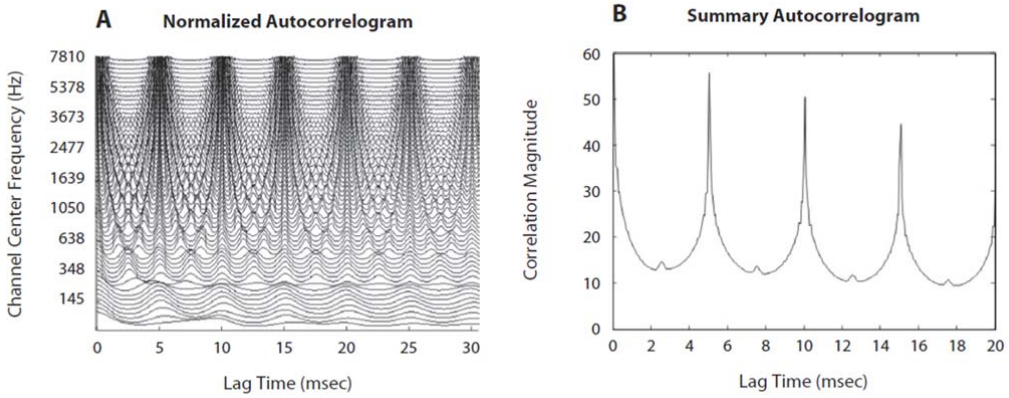

圖 1.10：自相關函數示意圖。[6]
左圖 A 為基頻為 200Hz 的複音在各頻段的自相關響應圖。
右圖 B 為將反應能量加總的自相關響應圖。

在 1997 年 Meddis 和 O'Mard，對於頻譜模型和時間模型做了總體評估[4]，將頻段依據可否解析的程度劃分為低(LOW:125-625Hz)、中(MID: 1375-1875Hz)、高(HIGH:3900-5400Hz)三個區段分別做討論，經過幾種實驗後他們發現到人在判斷音高時對於可解析的諧音(Resolved harmonics)與不可解析的諧音(Unresolved harmonics)的判斷分析方法有所不同。主要的區別為可解析的諧音對於音高的感知上有較強的影響，但在相位的感知上卻不強烈。反之，不可解析的諧音在音高的感知上的影響較弱，但相對的在相位的感知上很顯著。所以他們認為兩種音高感知模型皆有其道理，推斷人在音高的感知上並不只是單一的模型可以充分解釋，因而提出雙工音高感知模型(Unitary model of pitch perception)[7]。

然而這些模型皆是對於聽覺上的假設，目前仍無確切的證據證明符合這些聽覺模型生理構造的存在，只能從一些現象中推測而得。在本論文中，我們將嘗試運用類神經網路結合生理聽覺在音高感知上的模型，來抽取音樂的旋律並與其他方法的抽取結果做比較。

2. 提出系統架構及實驗結果

2.1 卷積神經網路簡介

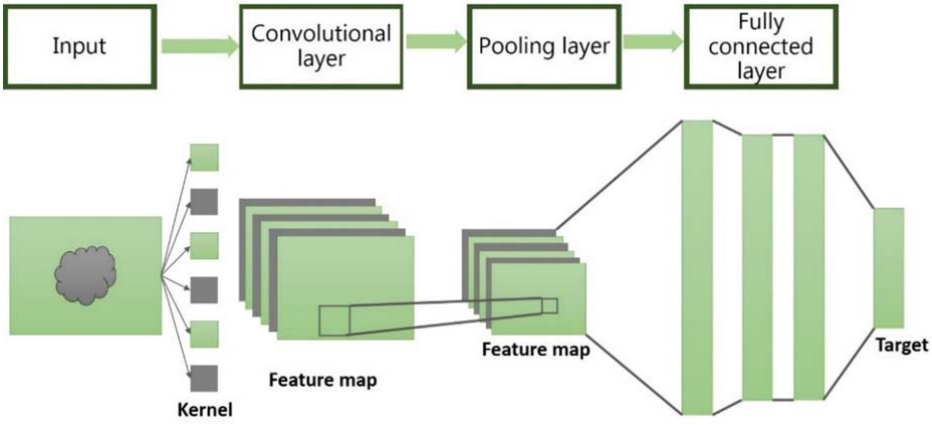

圖 2.1：卷積神經網路架構示意圖

卷積神經網路(Convolutional Neural Network, CNN)為神經網路的變形，其特色為神經元可以同時作用於一個區塊的資料，此網路在大型的影像處理應用上有出色的表現，一般的卷積神經網路包含三層：卷積層(convolutional layer)、池化層(pooling layer)以及特徵映射層(fully connected layer)。圖(2.1)為標準的卷積神經網路架構之範例。輸以是二維的原始圖像，在卷積層中經過與卷積核(kernel)的運算後，可以提取到其相對應的特徵圖(feature map)，每個卷積核所得到的特徵圖皆為一獨立平面，且其平面上所有神經元之權值相等，此步驟於物理意義上為提取與目標相關之特徵，以利我們之後的計算。圖(2.2)為卷積層數學運算之範例。

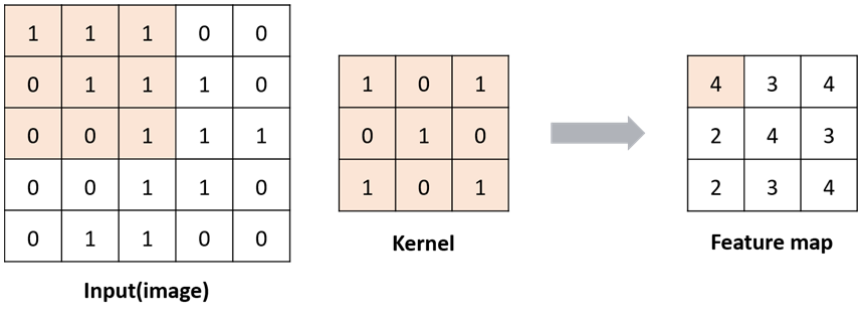

圖 2.2：卷積核大小為 3x3 之卷積層範例

因卷積神經網路適合對二維圖像資料進行分析，且具有權值共享、資料平移不變性的特點，非常符合音高的頻譜模型中使用模板找尋音高諧音所產生的固定特徵，因此我們在頻譜模型的實作上將採用卷積神經網路作為諧音比對偵測的架構。

首先我們討論諧音樣式在線性頻率頻譜圖及對數頻率頻譜圖上的差異，如圖(2.3)所示

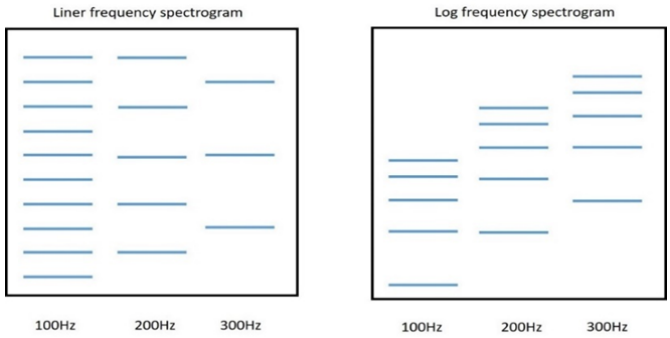

圖 2.3：不同音高的諧音在線性頻譜與對數頻譜比較圖

可由圖(4.8)看出若在頻率軸為線性的頻譜圖上，諧音的間距會隨著音高變高而變寬，而在對數頻率軸的頻譜圖上，不同音高諧音之間的間距並不會改變，我們認為人耳聽覺在能量及頻率上的感受符合對數關係，且在分析音高時有一定的形式，因此我們採用卷積神經網路作為訓練模型，並使用對數頻率軸的頻譜圖作為訓練輸入如圖(2.4)所示，可發現不同音高其諧音的結構在對數頻率頻譜圖上皆一致，僅是垂直平移的關係。我們將卷積核視為抓取諧音樣式的濾波器，預期卷積神經網路在做旋律音高辨識時與人的生心理聽覺特性相符。

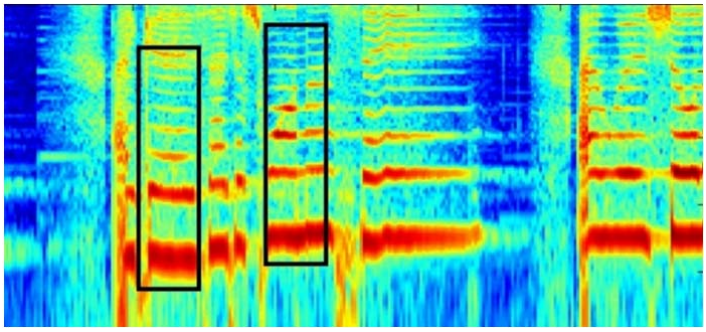

圖 2.4：範例樂曲片段之對數頻律軸頻譜圖

在文獻[8]中模擬神經彼此對於不同音高的交互反性，如圖(2.5)所示，認為人在音高判別上存在著一個呈對數分布的聽覺模板。

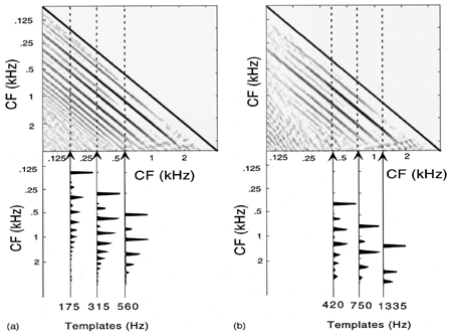

圖 2.5: 模擬神經對於不同音高的交互反映現象[8]

如圖(4.9)所示，我們認為在第一階段頻譜圖做卷積時應是做諧音的抓取，在對數頻譜圖上諧音的樣式(pattern)滿足位移不變性(shift-invariant)，文獻[8]中提出在人的聽覺系統中可能存在著一個抓取諧音的聽覺模板，因此我們認為在對數頻譜圖上只需要一個卷積核即可抓取到偵測音高重要的資訊。我們將依照此想法給定初始的卷積核樣式如下圖(2.6)。

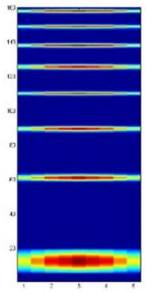

圖 2.6：初始卷積核樣式

依據我們對於頻譜圖的觀察，若要從卷積核看出諧音的樣式需要縱軸(頻率)方向包含的範圍夠大時才能觀察的出。因此我們設計卷積核為 5x160 的大小，並根據下式決定各個諧音間的位置。

$$h_i = A \times \log(f_0 \times i)$$

h_i 代表第 i 條諧音位置，A 和 f_0 則是調控整體起始位置及間距寬度在此分別設定為 48/log 和 1 使之寬度間距與輸入之對數頻譜圖相符，i 的範圍設定為 1 到 9，亦即僅選取 9 個諧音，因根據文獻[6]指出，人類在分析頻譜時超過第 10 個諧音將難以被解析出來。此外，音高對於時間的敏感度很高，我們推測輸入頻譜的中心時間資訊最為重要，所以對於每個諧音在時間軸上做高斯分佈；而每個諧音在頻率軸上會有散布的能量，因此我們對於每個諧音在頻率軸上的寬度也做了高斯分佈；在對數頻譜圖上，當頻率越低，諧音在頻率軸上的寬度也會越寬，依據此特性我們也將諧音在頻率軸上的寬度依據所在頻率做調整。

我們將探討隨機給定卷積神經網路的初始值(random initial) 以及依此模板(template)作為卷積核之初使值所訓練出來的結果，觀察兩者最終訓練出來對於旋律抽取的結果表現，以及卷積核所抓取的特徵是否具有物理意義，符合上述的假設得到預期的效果，從而能更加瞭解深度學習所習得的內容，得知抓取音高旋律時重要的特徵並簡化模型架構，以至於能針對旋律抽取設計出適合的架構。下圖(2.)為基於頻譜模型的旋律抽取架構流程圖。

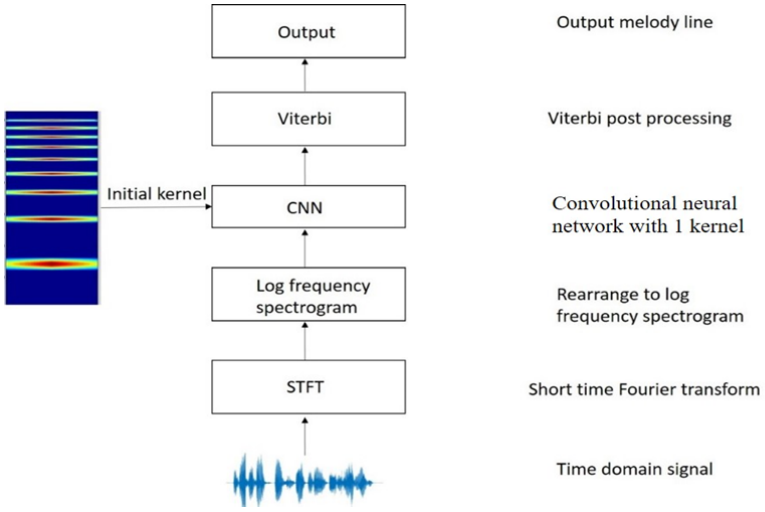

圖 2.7：基於頻譜模型之神經網路於旋律抽取之流程圖

首先我們將輸入的訊號降取樣至 16kHz 以減資料算量，經由短時傅立葉轉換(short time Fourier transform)計算出音訊的頻譜圖，因為人在聽覺上對於頻率及能量的感受皆呈對數關係，我們將計算出來的頻譜圖重新排列成對數頻譜圖(log-frequency spectrogram)，並將能量取對數運算變成對數能量(log-power)頻譜。經由卷積神經網路訓練出頻譜圖及音高的對應關係。通常訓練卷積神經網路會有多個卷積核可視為不同種類的濾波器，然而因我們認為不同音高的聲音在對數頻率軸的頻譜圖上僅是憑一的關係，因此在諧音萃取階段僅需一個濾波器就足夠，在 2.5 的實驗結果將會對此做更深入的探討。最後再針對神經網路輸出音高序列做維特比(Viterbi)解碼後處理得到較平滑的旋律線。

2.2 時間模型

計算自相關函數(Autocorrelation function)是一在音高偵測上很常被使用到的方法，至今仍有很多基於此方法的延伸改良，其主要的概念是因為有音高的聲音會在一定的時間內產生週期性的震盪，藉由計算自相關函數的方式可以突顯出震盪的週期，而理想中這些震盪周期的倒數即為音高頻率。然而在實際的情形目標訊號可能會受到噪音、空間效應及錄音裝置等影響，或是發聲的始末會有較難預估的暫態響應皆會影響自相關函數計算出來的結果，許多研究皆是針對這些問題提出不同的方法以進行改良。

在這裡我們保留原始計算出的自相關函數，使類神經網路針對旋律抽取的問題做學習建立出適當的模型，因為在時域訊號上並不會有高頻無法解析的問題，我們將探討在頻譜模型頻率無法解析的區段在時間模型上的表現，因此我們會將對輸入訊號做頻譜白化(frequency whitening)的預處理，藉此強調出高頻成份，流程架構如圖(2.8)。

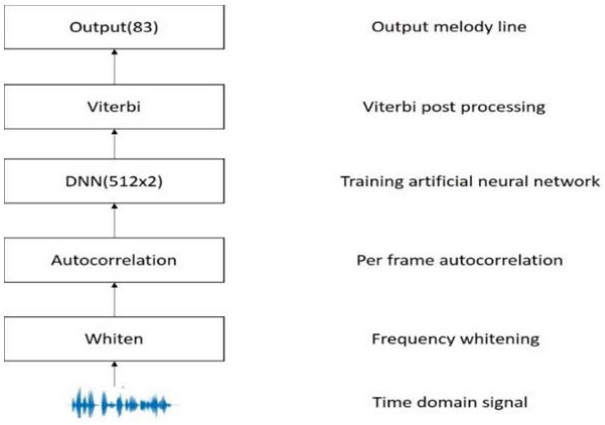

圖 2.8：基於時間模型之神經網路於旋律抽取之流程圖

2.3 雙工模型

在文獻[32-33]中提出，經由實驗數據推斷，人在分析音高時不僅只參考一種聽覺模型之輸出，而我們在上述研究發現確實在頻域及時域上對於旋律音高的偵測分別皆有效果，再加上在生心理聽覺上頻譜分析會有無法解析的區段，因此我們嘗試結合頻譜模型及時間模型兩個方法，期望藉由生理聽覺的概念出發，從不同面向的分析以及資訊的互補，建構出基於雙工模型之類神經網路的音高判別演算法，架構如圖(2.9)。

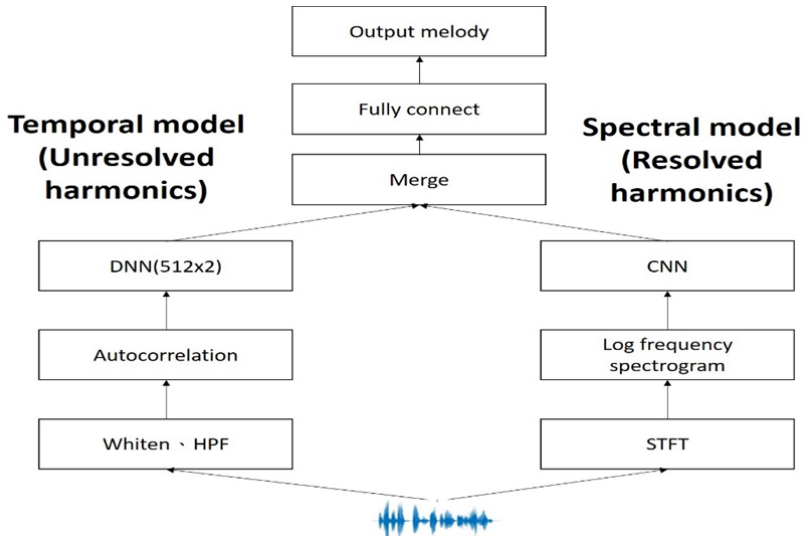

圖 2.9：基於雙工模型之神經網路於旋律抽取之流程圖

整體架構主要上述兩種模型作合併，為了區分出兩個模型各自的效用，我們在雙工模型中的時域部分的輸入多加了高通濾波器的預處理，使時域部分特別針對頻域無法解析的部分做補足，而高通濾波器的截止頻率(cutoff frequency)的選擇我們參考文獻[4]對於頻率上可否解析諧音的分段我們選擇使用截止頻率為 1375Hz 的高通濾波器做預處理，經過時間及頻譜模型後再做整合的模型訓練。

2.4 維特比解碼

觀察神經網路的輸出值，我們發現有些時候會有可能會因為短時干擾而使預估的輸出值會有不符合預期的跳動，然而一般而言人所認知的旋律並不會在短時間內產生快速的變化，因此在神經網路的輸出之後我們嘗試使用維特比解碼做為後處理。我們以訓練資料的音高轉移作為依據，計算狀態轉移機率來做預估判斷，嘗試修正不符合預期的瞬間跳動值，得到較合理平滑的旋律線。依據訓練資料將所對應的音高做量化視為不同的狀態，計算狀態間的轉移次數產生狀態轉移機率矩陣(transition matrix)，依照此機率作為修正懲罰權重，對深度學習做出來的結果作修正。

2.5 評量方式

		Detected		
		unvx	**vx**	Sum
Ground truth	**unvoiced**	**TN**	**FP**	GU
	voiced	**FN**	**TP**	GV
	sum	DU	DV	TO

表 2.1 :旋律抽取評量方式

1.回答率(voicing recall rate)

經由演算法所估計出有聲音的音訊片段佔資料庫所標記答案有聲音的音訊片段之比例，與偵測理論中的擊中率相同(hit rate)。

$$Voicing\ recall\ rate = TP / GV$$

2.誤答率(voicing false alarm rate)

經由演算法估算出有聲音的音訊片段但在資料庫卻標記為無聲段落之比例。

$$Voicing\ false\ alarm\ rate = FP / GU$$

3.音高正確率(raw pitch accuracy)

計算在資料庫標記有音高的時間點經由演算法估計出的音高的正確率。在此計算音高的單位為赫茲(Hz)，且對於計算誤差有 1/2 個半音(50 音分)的容忍度。

$$Raw\ pitch\ accuracy = \left(TPC\ +\ FNC\right) / GV$$

4.音名正確率(raw chroma accuracy)

計算在資料庫標記有音高的時間點經由演算法估計出的音名的正確率，忽略八度音的估計誤差。

$$Raw\ chroma\ accuracy = \left(TPCch\ +\ FNCch\right) / GV$$

5.總體正確率(overall accuracy)

同時考量有無旋律的判別及音高偵測的結果，計算總體正確率。

$$Overall\ accuracy = \left(TPC\ +\ TN\right) / TO$$

2.6 實驗資料

本論文主要針對人聲旋律做預估，使用了兩種資料庫來進行演算法效能的評估，分別為 MIR1K 及 iKala 資料庫，實作上為了統一訓練資料格式並降低運算量，我們將 iKala 資料庫的音檔取樣率降至 16kHz 使之與 MIR1K 相同，又因為兩個資料庫皆是設計給歌聲分離使用的雙聲道音檔，在本論文的實驗中我們皆以訊噪比(signal to noise ratio)為 0dB 的方式將歌聲與背景音樂做混合，成為單聲道的音檔。iKala 資料庫我們將前 200 首作為訓練資料，其餘 52 首為測試資料。MIR-1K 資料庫我們取前 720 個音樂片段作為訓練資料，後 280 個音樂片段為測試資料。

2.7 實驗設定

首先我們先設定欲偵測的音高區間，根據[9]的實驗設定考量合理的人聲音高範圍，我們將音高從 D2(73.4Hz)到 F#5(740Hz)，按照每 50 音分做頻率的量化(quantize)成為 82 種狀態如圖(5.1)所示，將音高偵測視為分類問題。而我們希望系統預期能同時達成音高偵測及有無旋律的判別因此再增加一個無聲(unvoiced)的狀態，總共將有 83 種狀態輸出。

我們先將音檔取樣頻率降至 16kHz，針對頻譜模型的輸入使用 1024 點的短時傅立葉轉換來做頻譜分析，音框大小為 48ms，音框重疊率為 50%，再將頻譜重新排列成每八度 48 點 (48 points per octave)的對數頻譜圖。觀察人的歌聲頻率分布及對數頻譜的解析程度，最後選用 250 點(約 5 個八度)的對數頻譜圖，觀察範圍約為 100 到 3700Hz。由文獻[9]指出考量前後音高的資訊對於判別當下音高有幫助，因此我們擴增當下時間前後各 3 個音框，最終產生維度為 7x250 的小區域頻譜圖作為輸入，如圖(2.10)所示。

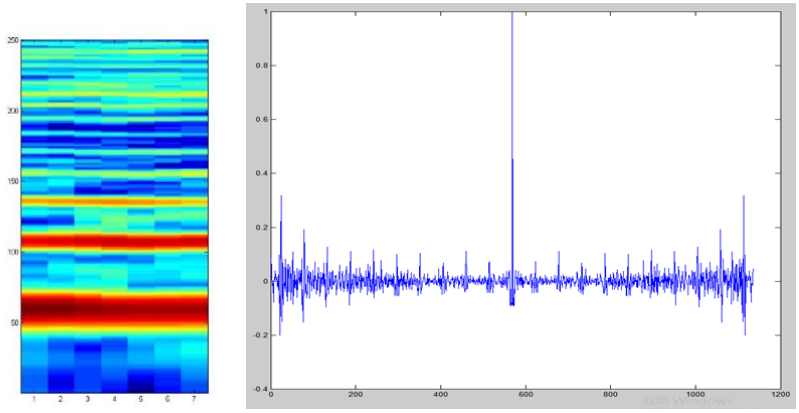

圖 2.10：頻譜模型輸入資料

針對時間模型，我們採用自相關函數作為輸入，首先將音檔取樣頻率降為 16kHz，再經由 10 階的巴特沃斯高通濾波器，截止頻率為 1375Hz，音框大小與重疊率皆與考慮頻譜模型時的設定相同，一個音框大小為 48ms，也就是 768 個取樣點，我們先對音框內的頻率進行白化，如圖(5.4)，將高頻部分強調出來，經由自相關運算後得到 1535 個點，因為邊界值的計算誤差較大，我們最後取鄰近中心點的自相關函數 1136 維的資料作為時間模型的訓練輸入。

2.8 實驗結果

2.8.1 頻譜模型

首先我們先將問題簡化，使用純歌聲無背景音樂作為訓練目標，觀察卷積神經網路訓練的結果。實作上我們卷積神經網路的架構為一層卷積層(convolutional layer)及兩層的 512 點的(fully connect layer)及 83 維的輸出層(output layer)。

當不給定神經網路卷積核初始值，經訓練後所得的卷積核樣式為圖(2.11A)，可發現其樣式類似於對數分佈，代表第一層卷積層確實是在做諧音的抓取，與我們的假設相符。我們進一步的觀察卷積層的輸出，其樣式為圖(2.11B、C)。為了方便觀察整體輸出樣式，我們將每一個資料經由卷積層的輸出值只取出時間中心的資訊(第 4 欄)，將多個輸出資料整併成一個段落，而該段落所對應的參考旋律標記為圖(2.13)。由圖中我們可以發現到在第一層的卷積層的輸出其實已大致描繪出音高走勢，其樣式也與參考標記極為相似。但在圖(2.12)中也可以明顯看出同一個時間約有兩條較明顯的輸出值，發現皆呈約 48 個單位的間距，而我們所使用的對數頻譜圖的設定正好就是每八度 48 點(48 points per octave)，因此我們合理推測其原因為諧音所造成的八度誤差(octave error)。在偵測音高旋律的議題上八度誤差是很常遇到的問題，傳統的方式多是觀測輸出的音列，給定一些準則加以限制並修正，而在此我們認為卷積層之後所連接的特徵映射層(fully connected layer)有助於解決此問題，可由評量結果數據看出。

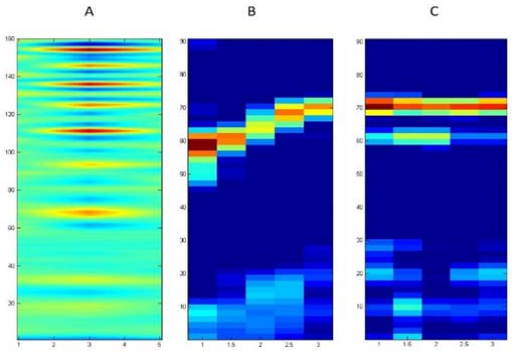

圖 2.11：隨機初始訓練結果。A 為卷積核，B、C 為某例卷機神經網路輸出

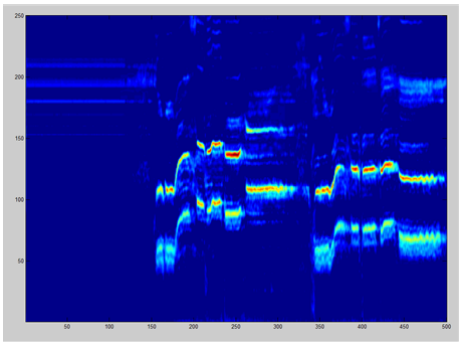

圖 2.12：第一層 CNN 之輸出樣式(某範例音樂片段)

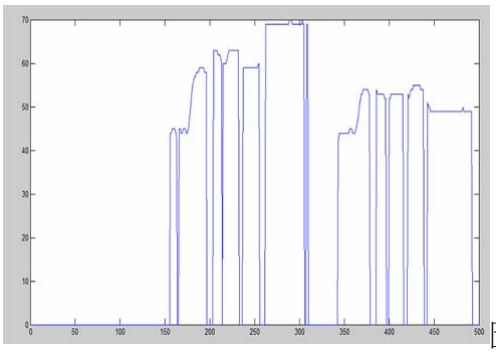

圖 2.13：圖(2.12)片段的參考音高標記

2.8.2 討論不同個數的卷積核的結果

一般而言，訓練卷積神經網路會用多個卷積核萃取不同特徵，上述實驗中因為我們依據生心理聽覺的假設只使用一個卷積核，在這裡我們將探討使用不同個數的卷積核是否對結果有所影響，為了專注探討此議題，在此我們皆不連接特徵映射層作訓練。以下分別列出卷積核樣式、卷積層輸出及輸出分數評量圖(2.14)、表(2.2)。

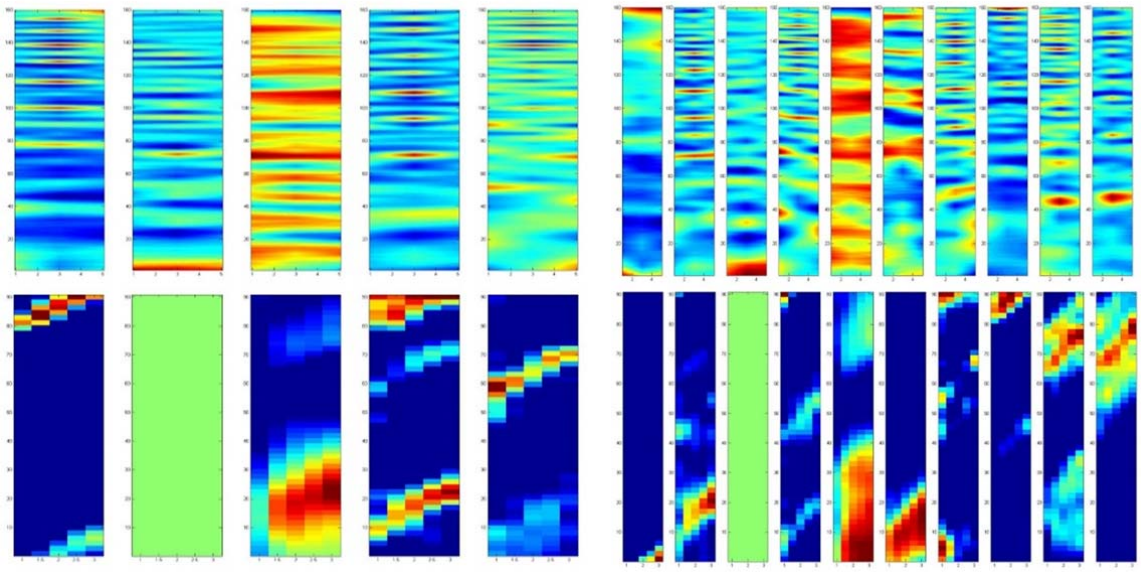

圖 2.14：5、10 個卷積核訓練結果

Kernel number	5	10	15
Recall	91.42%	90.38%	92.09%
False alarm	11.55%	9.78%	10.95%
Raw pitch	87.31%	86.48%	87.64%
Raw chroma	88.05%	87.42%	88.36%
Overall accuracy	88.60%	88.68%	89.05%

表 2.2：不同個數卷積核結果比較

客觀分數評量可以看出增卷對於積核的個數對於效能的影響並不顯著，原因推測如上述所提，增加卷積核並沒有顯著的增加資訊豐富度。且增加卷積核個數也是在增加模型的複雜度，計算的時間也相對提高。以上這些現象皆支持對於音高諧音的抓取僅需 1 個模板即足夠，且卷積神經網路的訓練結果與生心理聽覺感知中的頻譜模型概念非常相似，可以用一個聽覺模版來找出對應音高。

如圖(2.15)所示，給定卷積核初始模板再做訓練，我們發現卷積核的樣式確實會跟照著給定的模板樣式作微調，但是經由多次的訓練後抓取基頻的諧音線會漸漸消失，推測這個現象是因為對數頻率軸對低頻的解析度很高，因此低頻的聲音在頻譜圖上會看起來比較擴散而模糊，所以資訊價值較小。

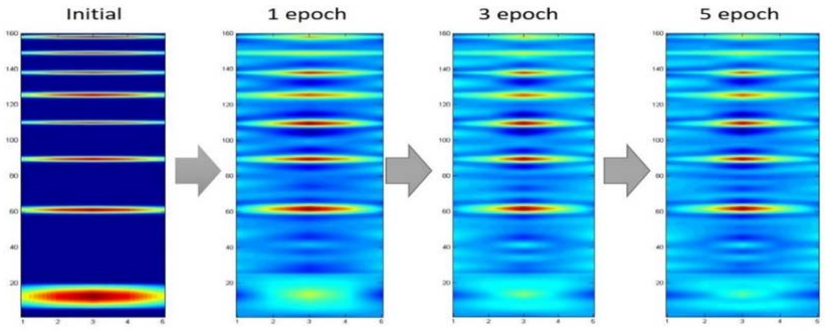

圖 2.15：給定模板後訓練趨勢

我們觀察出由卷積神經網路所得的輸出值有不連續的跳動，因此我們使用同樣的訓練資料集的音高標記計算狀態轉移矩陣，基於此矩陣做維特比解碼的後處理來修正一些不連續的狀態。下圖顯示針對某範例音樂片段，參考標記資料、原始輸出及經由維特比解碼修正後的結果。可觀察到經由維特比解碼的後處理可以修正一些不連續的跳動，得到較平滑的旋律線如圖(2.16)所示。加入維特比解碼後之整體效能評估分數如表(2.3)。

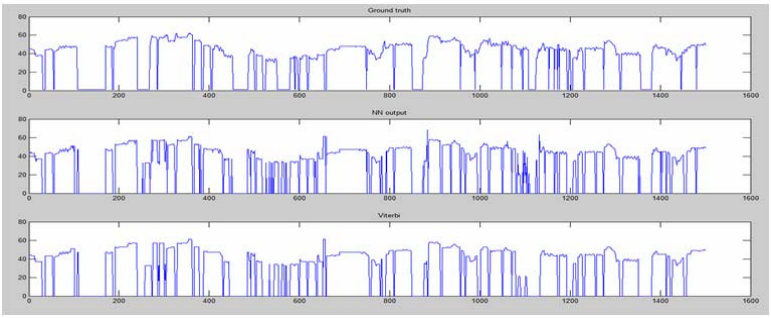

圖 2.16：維特比解碼後處理的輸出比較(某範例音樂片段)。
(上)原始參考標記、(中)類神經網路輸出、(下)維特比解碼輸出

	Without Viterbi	With Viterbi
Recall	84.78%	85.44%
False alarm	15.11%	15.51%
Raw pitch	75.43%	76.40%
Raw chroma	77.03%	78.22%
Overall accuracy	78.55%	79.07%

表 2.3：維特比解碼結果比較

2.8.3 時間模型

在頻譜模型上我們已驗證可解析的諧音對於旋律音高的判別有不錯的效果。因此在時間模型上我們首先要驗證在頻譜模型不可解析的頻段，加上時間軸上的資訊是否對於判斷旋律音高有所幫助。與之前實驗設定所述相同，我們使用混和人聲與背景音樂的時域訊號自相關函數作為訓練輸入，比較不同頻段輸入對於旋律抽取的效能差異，結果於表(2.4)。

	All pass	HPF(1375Hz)
Recall	78.12%	75.82%
False alarm	29.14%	28.52%
Raw pitch	69.94%	62.12%
Raw chroma	73.32%	65.31%
Overall accuracy	70.25%	63.56%

表 2.4：iKala 資料庫混合背景音樂在時間模型經高通濾波器之結果比較

由結果看出經由高通濾波器後的訊號仍有一定的效能，此現象也說明了在頻譜模型無法解析的範圍仍存在著對於旋律音高判別有價值的資訊。然而，時域訊號上的資訊較為雜亂，也較易受到雜訊的干擾，因此採用時間模型整體的效能較於採用頻譜模型時來得差是可預期的。

雙工模型

我們觀察到在頻譜模型無法解析的頻段上的訊號在時間模型上仍產生一定的效果，因此我們將兩者模型作合併，希望能同時考量可解析與不可解析的頻段，達到互補的效果。雙工模型與個別頻率、時間模型的效能比較於表(2.5、2.6)。

iKala	Spectal	Temporal	Hybrid
Recall	84.78%	75.82%	89.26%
False alarm	15.11%	28.52%	18.94%
Raw pitch	75.43%	62.12%	80.11%
Raw chroma	77.03%	65.31%	81.60%
Overall accuracy	78.55%	63.56%	80.42%

表 2.5：iKala 資料庫在雙工模性下的評分比較

MIR-1K	Spectral	Temporal	Hybird
Recall	83.63%	81.57%	82.73%
False alarm	21.31%	26.76%	16.14%
Raw pitch	68.81%	67.87%	72.23%
Raw chroma	72.16%	71.71%	75.38%
Overall accuracy	71.70%	69.44%	75.64%

表 2.6：MIR-1K 資料庫在雙工模性下的評分比較

由表可看出利用類神經網路綜合考量時間與頻率的資訊確實對於旋律抽取的效能有所提

升，結合兩者從不同面向的觀測確實有互補的效用。

最後我們與其他旋律抽取的方法做比較，為了公平性我們皆使用相同的訓練及測試資料集做衡量，其中 MCDNN 的方法依循文獻[9]的描述做出，輸入資料則與所提出的頻譜模型相同。我們與近期旋律抽取的議題上著名的專家系統 Melodia 比較[10]，並考量文獻[11]使用 HPSS(Harmonic/Percussive Source separation)[12]經由聲源分離前處理的訓練結果呈現於表(2.7、2.8)。

Algorithm	Proposed	HPSS	MCDNN	Melodia
Recall	89.26%	83.42%	85.85%	82.02%
False alarm	18.94%	13.92%	15.05%	26.71%
Raw pitch	80.11%	74.43%	77.88%	75.99%
Raw chroma	81.60%	75.97%	79.60%	78.36%
Overall accuracy	80.42%	78.28%	80.22%	72.80%

表 2.7：使用 iKala 資料庫混和背景音樂與其他系統比較

Algorithm	Proposed	HPSS	MCDNN	Melodia
Recall	82.73%	75.35%	78.36%	85.10%
False alarm	16.14%	12.37%	14.25%	30.80%
Raw pitch	72.23%	64.29%	65.21%	72.95%
Raw chroma	75.38%	67.72%	68.30%	75.74%
Overall accuracy	75.64%	71.12%	71.22%	69.61%

表 2.8：使用 MIR-1K 資料庫混和背景音樂與其他系統比較

3. 結論

本文利用類神經網路架構模擬人的生心理聽覺提出旋律抽取演算法。基於過往生心理聽覺學者在人類音高感知上的研究，根據頻率的解析與否主要分為頻譜及時間兩種聽覺模型，我們使用類神經網路實現出這些生心理聽覺的音高感知模型，來抽取音樂片段中的旋律。頻譜模型上我們使用卷積神經網路，因其架構的特性符合模板濾波器的觀點，經實驗發現存在一個抓取音高的聽覺模板，即可達到對所有諧音音高的識別，增加卷積核的個數對於旋律抽取的效能並無顯著的提升，因此我們也提出了聽覺感知模板作為訓練初始值增進訓練的收斂速度。時間模型我們使用自相關函數作為輸入資料，我們發現在頻譜模型無法解析之頻律處在時間模型中仍有效果，這現象也符合偵測殘餘音高(residual pitch)的理論。因此我們使用時間模型對於頻率模型頻率不可解析處做補強，建立出結合頻譜及時間資訊的雙工模型，運用神經網路的整合達到互補以提升效能。

針對頻譜模型的輸入訊號我們嘗試使用不同的輸入資料如諧音與敲擊聲的分離(HPSS)，以及維特比解碼的後處理，兩者對於訓練結果並沒有顯著的提升，代表類神經網路能從原始資料(raw data)中找出適合旋律抽取的特徵及方法。此外我們也嘗試將背景噪音改為混和白雜訊或是工廠噪音等非諧音噪音時表現明顯較好，因此在非諧音噪音干擾下有不錯的表現。

參考資料

[1] 張斌. *耳鼻喉科學*, 正中書局, 台北 (1996).

[2] Jing Chen, Thomas Baer, and Brian CJ Moore, "Effect of enhancement of spectral changes on speech intelligibility and clarity preferences for the hearing impaired," *J. Acoust. Soc. Am.,* 131.4: 2987-2998, 2012.

[3] Takashi Yamauchi, Presentation on theme: "Sensation & Perception", http://slideplayer.com/slide/6639448/

[4] Ray Meddis and Lowel O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Am.*, 102.3: 1811-1820, 1997

[5] Roy D. Patterson, Mike H. Allerhand, and Christian Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.*, 98.4: 1890-1894, 1995.

[6] William A. Yost, "Pitch perception," *Attention, Perception, & Psychophysics*, 71.8: 1701-1715, 2009.

[7] Robert P. Carlyon, Comments on "A unitary model of pitch perception" [J. Acoust. Soc. Am. 102, 1811–1820 (1997)], *J. Acoust. Soc. Am.,* 104, 1118 ,1998.

[8] Shihab Shamma, and David Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *J. Acoust. Soc. Am.*, 107.5: 2631-2644, 2000.

[9] Sangeun Kum, Changheun OH, and Juhan Nam, "Melody Extraction on Vocal Segments Using Multi-Column Deep Neural Networks," In *Proc. of ISMIR*, pp. 819-825, 2016.

[10] Justin Salamon, and Emilia GÓMEZ, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. on Audio, Speech, and Language Processing*, 20.6: 1759-1770, 2012.

[11] François Rigaud, and Mathieu Radenen. "Singing Voice Melody Transcription Using Deep Neural Networks," In *Proc. of ISMIR*, 2016.

[12] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," *2008 16th European Signal Processing Conference*, Lausanne, 2008, pp. 1-4.